Low Power TCAM Design And Simulation

Rahul Nigam
Department of electronics and communication, NIT, Calicut- India.

Abstract

This paper presents the approach to reduce power consumption in a ternary content addressable memory (TCAM). The main challenge with the TCAM design is to reduce the power consumption without sacrificing speed and area. Here in this paper I am doing practical implementations of a TCAM oriented for low-power applications. Low power TCAM designs have done 0.18µm CMOS technology.

Keywords: *Content addressable memory (CAM), feedback circuits, high speed, low power, matchline sense amplifiers (MLSA).*

1. Introduction

Content addressable memory (CAM) provides a fast data-search function by accessing data by its content rather than its memory location indicated by an address. In addition to the conventional READ and WRITE operations, CAMs also support SEARCH operations as compared to RAM. CAM allows searching its entire contents within a single clock cycle, i.e. parallel lookup capability. CAM can be used in a wide variety of applications such as parametric curve extraction, Hough transformation, Huffman coding/decoding, Lempel–Ziv compression and image coding, data base access, pattern matching and networking IP address lookup etc. Now a day main application of TCAM is to classify and forward IP packets in network routers. TCAM is required to implement the masking function i.e. storing X (don't care) in the TCAM cell. [1][2][3]

On the other hand major disadvantages of the CAM are the high power dissipation and the high area cost.

2. CAM Cell Types

2.1. Binary Cells

Here 10 T binary CAM is shown in **Figure** 1. A CAM cell consists of two basic components: storage element and comparison logic. The storage element has implemented with a SRAM cell and the comparison logic usually executes XNOR function. Transistors N1-N4 implements XNOR logic function.[8] [10]

2.1.1. Write Operation

The write operation is done by placing the data on the bit line and enables the word line. This turns on the access transistors (N6-N7) and the internal nodes of the inverters are store the BL data.

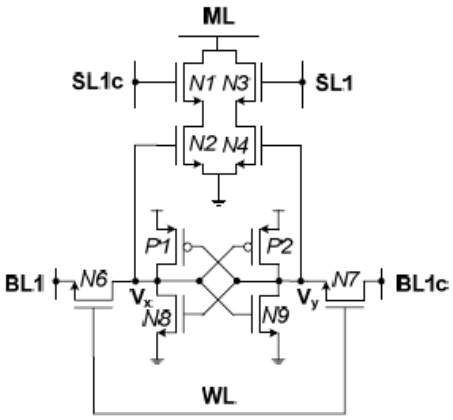

Figure 110-T Binary CAM

Assume V_X = 0 and V_Y = 1 P2 and N8 were 'ON' and P1 and N9 were 'OFF' and we want to WRITE 1 in cell. For WRITE 1 we put BL1='1' and BL1c='0' and when wordline is enabled (WL = '1') access transistors (N6-N7) conduct resulting in BL currents. To overpower the feedback inverter we need access transistors larger size as compare to P1 and P2. [1][4]

2.1.2. Read Operation

The READ operation is done by precharging the BL1 and BL1c to V_{DD} and enables the word line (WL). If V_X = 1 and V_Y = 0 then current I_{READ} discharges BL1c (through N7 and N9). BL1 remains at V_{DD} because V_X = '1'. Therefore a small voltage difference develops between BL1 and BL1c. The current I_{READ} raises the voltage V_X. Thus the driver transistors (N8-N9) are sized such that V_X remains below the inverter threshold voltage and hence the cell does not write at the time of the READ operation. Typically the driver transistors (N8-N9) are sized 1.5 times wider than the access transistors (N6-N7). [1][4]

2.1.3. Search Operation

The SEARCH operation is done in three steps. First we precharge search lines (SLs) SL1 and SL1c to GND. Then ML is precharged to V_{DD}. Finally the search data bits placed on searchlines SL1 and SL1c. If the search data bit is identical to the stored value (SL1=BL1, SL1c=BL1c) both ML to GND pull-down paths remain 'OFF' and the ML remains at V_{DD} indicating a 'match'. Otherwise one of the pull-down paths conducts and discharges the ML to GND indicating a 'mismatch'. Precharging SL1 and SL1c to GND during the ML precharge ensures that both pulldown paths are 'OFF'. [1][4]

2.2 TCAM Cells

A typical 16T static TCAM cell is shown in **Figure** 3. It is similar to the binary CAM cell except that it has two SRAM cells to store ternary data. READ, WRITE and SEARCH operations in this cell are done in the same way as described earlier. For the masking we need to be turn off both ML to GND pulldown paths. For example global masking is done by SL1 = SL2 = '0' and local masking is done by $V_X = V_Y$ = '0'. [1][4]

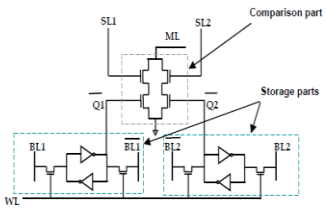

Figure 3 16-T TCAM Cell

2.3 CAM Array

A CAM word of n bit is implemented by connecting n CAM cells in parallel. All the cells in a CAM word share an ML but they have separate SLs. The ML is connected to a ML sense amplifier (MLSA) which determines word matches with the search bits or not. During search operation the ML remains at V_{DD} only if all the bits result in 'match'. In other words even if a single bit 'mismatch' result in a discharge path for ML and indicating a word 'mismatch'. A CAM array (m*n) is implemented by m CAM words with the same set of SLs. The search bits (n bits) are written on SLs is compared with all the m words in parallel.[9]

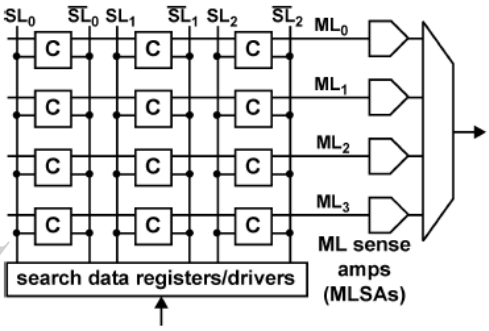

Figure 4 TCAM Array

3. Matchline Sense Amplifier

3.1 Conventional MLSA

Initially precharge all the MLs to V_{DD} and the search bits are applied on the SLs. If a TCAM word is identical to the search bits then the ML remains at V_{DD}. Otherwise it discharges to GND. In order to avoid a short circuit current the SLs are precharged to GND during the ML precharge phase. Hence most of the SLs switch in every SEARCH operation causing high power consumption.

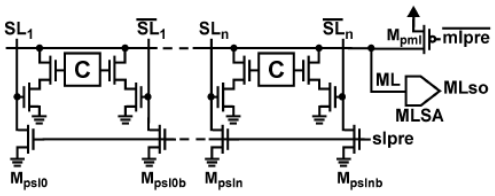

Figure 5 Conventional Precharge MLSA

3.2 Current Race Sensing Scheme

To reduce power consumption during SEARCH operation we use this CR scheme. In this scheme the ML precharge to GND during precharge phase so there is no need to precharge searchline SLs. Thus the average SL switching activity can be reduced approximately by half. The ML sensing is initiated by charging up the ML using a constant current source. Since a match word does not have a current discharge path it charges at a faster rate than a mismatching ML. In the matching condition when ML charge above the NMOS threshold voltage V_{TN} its MLSO changes from '0' to '1' (as shown in **Figure** 6).

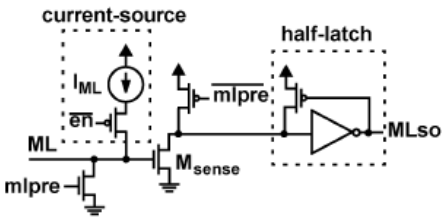

Figure 6 Current Race MLSA

The ML capacitance can be given by equation

$$C_{ML} = [2g+4(n-g)]\, C_{DRAIN} + C_{INT} + C_{MLSA}$$

Where 'g' is the number of globally masked bits, n is the total number of bits per word, C_{DRAIN} is the drain capacitance of each transistor in the comparison logic, C_{INT} is the interconnect capacitance of each ML and C_{MLSA} is the MLSA input capacitance. Like the first term in equation C_{INT} is also proportional to n. However for large values of n C_{MLSA} is negligible as compared to the first two terms. When a bit is globally masked (SL1 = SL2 = '0') only the drain capacitances of transistors N1 and N3 (shown in **Figure** 1) contribute to C_{ML}. Otherwise CML also includes the capacitance of the internal nodes. Therefore the worst case C_{ML} corresponds to no global masking g = 0 and the best case C_{ML} relates to full global masking i.e. g = n. [1][6]

3.3 MLSA With Resistive Shielding

Here it uses an NMOS transistor (N3) in the triode region to decouple the ML and its MLSA. The N3 channel resistance shields the sensing point SP from the highly capacitive ML. Due to the body effect and the decreasing gate to source voltage the N3 channel resistance increases when the ML voltage is rising up. The N3 channel resistance depends on the number of mismatch bits. For instance ML_0 would be rising faster than ML_1 (ML_0 match word, ML_1 1 bit miss) which implies that N3 of ML_0 has higher resistance to shield the node SP from the ML. Since less current is now being diverted to the ML, the node SP charges much faster to reach the threshold voltage. Faster sensing of the ML_0 also reduces energy consumption because the ML current sources are shut down sooner.[3][5]

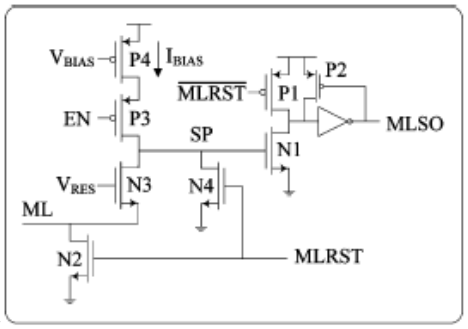

Figure 7 MLSA with Resistive Shielding

3.4 MLSA with Active Feedback

Here transistor N3 operates as a constant current source (I_{FB}). The MLEN signal enables the MLSA by activating EN, I_{BIAS} and I_{FB}. Initially all ML_s receive the same current from the current sources I_{BIAS}. As ML_0 charges at a faster rate than ML_k its P6 source to gate voltage becomes smaller than that of ML_k. In order to keep the current through P6 constant I_{FB} a reduction in V_{gsp6} is compensated by an increase in P6 source to drain voltage V_{sdp6}. Since the source terminal of P6 is close to V_{DD} (P7 is acting as a switch) a larger V_{sdp6} results in a smaller V_{cs}. Thus the faster charging of ML_0 makes its V_{cs0} smaller than that of V_{csk}. As a consequence ML_0 receives higher current and charges more rapidly than ML_k. This positive-feedback action continues until ML_0 reaches the MLSA threshold voltage and switches MLSO 0 to 1 which result in turns off the current sources by switching EN 0 to 1.[5][7]

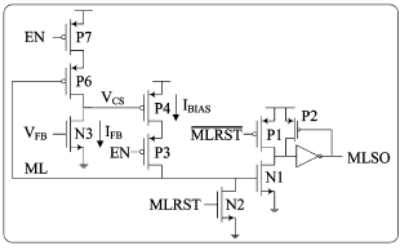

Figure 8 MLSA with Active Feedback

4. CIRCUIT DESIGN

4.1 TCAM Cell

Each TCAM cell contains two SRAM cells. The SRAM area was minimized by choosing minimum size transistors 0.42/0.18 wherever possible. The cells have designed to perform the READ operation as well thus the driver transistors (N8/N9) were sized 1.5 times larger than the access transistors (N6/N7) as shown in **Figure** 1

4.2 CR MLSA

In the conventional CR MLSA the ML current source I_{BIAS} was implemented using large size PMOS transistors to support a current that is high enough to match the speed of the positive feedback MLSA. A weak transistor PMOS was included to compensate for M_{SENSE} leakage while holding the node MLSO at '0'. Transistor M_{SENSE} was sized relatively large to override PMOS as shown in **Figure**6.

5. SIMULATION AND MEASUREMENT RESULTS

5.1 TCAM cell Operations

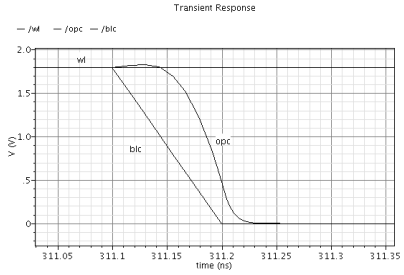

Figure 9 WRITE 0 in TCAM cell

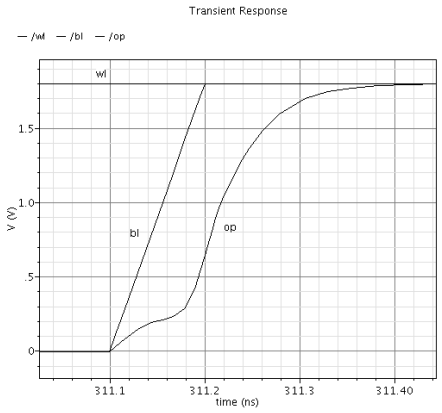

Figure 10 WRITE 1 in TCAM cell

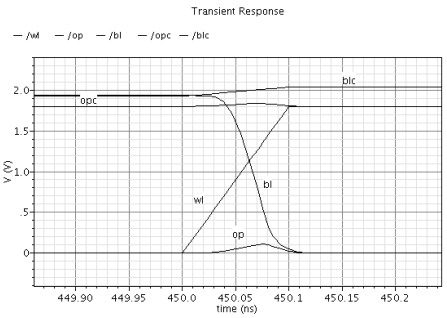

Figure 11 READ operation in TCAM cell

Here OPC (output complement) is high and OP (output) is low, so when WL is enable, the bit line which is precharged to high initially, reduces its value. After this bit line sense amplifier (BLSA) sense the difference between two bit lines.

Table 1 READ and WRITE operation

Operation	Write 1	Write 0	Read
Delay(ns)	0.0618	0.0392	0.0184
Energy(fJ)	19.4	4.13	2.73

TCAM ARRAY (16 *16)

Table 2 STORED DATA

1011	1000	1001	1100	WORD1
1001	1000	1001	1100	WORD2

1000	1000	1001	1100	WORD3
xx11	1000	1001	1100	WORD4
1100	1000	1001	1100	WORD5
1011	0111	1001	1100	WORD6
1001	1001	1010	1101	WORD7
1000	1001	1010	1101	WORD8
x011	1000	1001	1100	WORD9
xx11	1000	1001	110x	WORD10
xx11	1000	1001	11xx	WORD11
1000	1001	1010	1111	WORD12
1000	1011	1010	1111	WORD13
0000	1011	1010	1111	WORD14
0000	1111	1010	1111	WORD15
0000	0000	1010	1111	WORD16
SEARCH KEY	1011	1000	1001	1100

5.2 Current race MLSA

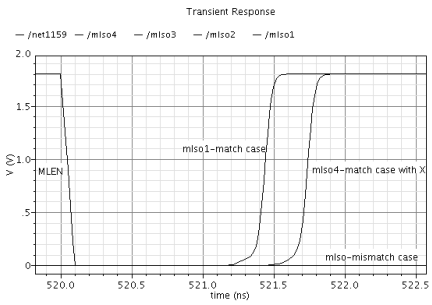

Figure 12 Current race MLSA output

If we mask the bits (locally) then delay will increase as shown in **Figure** 12.

Search time non masking case is 1.3855 ns

Masking case 2 bits 1.6853 ns

5.3 Resistive feedback MLSA

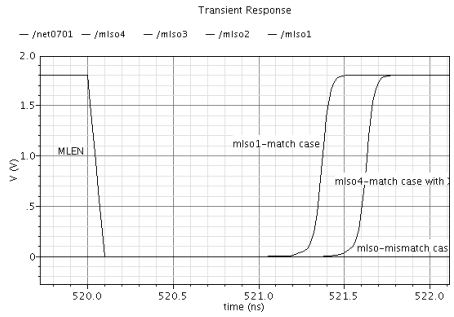

Figure 13 Resistive feedback MLSA output

Search time non masking case is 1.3178 ns

Masking case 2 bits 1.5769 ns

5.4 Active feedback MLSA

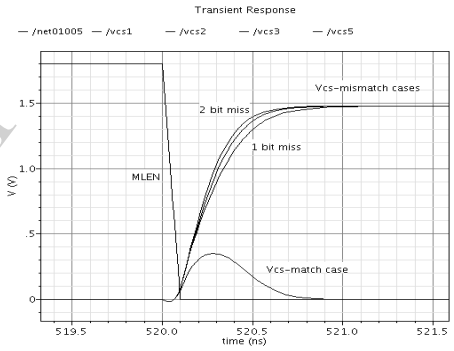

Figure 14 Voltages at VCS_k node

Vcs value in the match case is small as compare to the V_{csk} (k bits mismatch). So it provides more current to ML as discussed earlier.

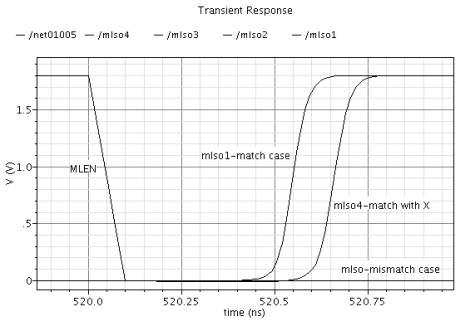

Figure 15 Active feedback MLSA output

Search time non masking case is 0.4992 ns

Masking case 2 bits 0.5995 ns

Table 3 Comparison between different MLSA

Scheme	Search time (ns)	ML Energy (FJ)	
		Match case	Mismatch case
Current Race	1.3855	64.88	2.26
Resistive Feedback	1.3178	32.80	2.11
Active Feedback	0.4992	26.74	4.76

7. CONCLUSION

This paper uses the four types of matchline sense amplifiers to reduce the power consumption in search operation and increase speed. In CR MLSA we have no need to precharge the searchline and also because of ML precharge to low so there is no charge sharing problem as compared to conventional scheme. In this we charge all the matchline with same amount of current. Ideally MLSA should provide maximum current to ML in case of match (for increase speed) and minimum amount of current for mismatch case (to reduce power consumption). If we apply positive feedback in MLSA then it provides more current for the match case as compared to mismatch case. Here in this project I have used two types positive feedback MLSA which combines both i.e. reduce power consumption and increase speed.

8. REFERENCES

[1] Kostas Pagiamtzis, Student Member, IEEE, and Ali Sheikholeslami, Senior Member, *IEEE "Content-Addressable Memory (CAM) Circuits and Architectures: A Tutorial and Survey"* IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 41, NO. 3 MARCH 2006.

[2] Konstantinos Masselos *"Lecture 7Memory and Array Circuits"* Department of Electrical & Electronic Engineering Imperial College London URL : http : // cas . ee.ic.ac.uk Email: k.masselos@ic.ac.uk

[3] M Sultan M Siddiqui and G .S Visweswaran *"A High Performance and Low Power TCAM Solution for Packet Forwarding in communication"*, Electrical Engineering Departmen Indian Institute of technology Delhi New *Delhi* NCC 2009, January 16-18, IIT Guwahati.

[4] Saleh Abdel-Hafeez1, Shadi M. Harb2, and William R. Eisenstadt2 *"Low-Power Content Addressable Memory With Read/Write and Matched Mask Ports"* Department of Computer Engineering, Jordan University of Science & Technology Irbid, Jordan 21110 sabdel@just.edu.jo Department of Electrical & Computer Engineering University of Florida, Gainesville, FL 32611.

[5] .Nitin Mohan, Member, IEEE, Wilson Fung, Member, IEEE, Derek Wright, Student Member, IEEE, and Manoj Sachdev, Senior Member, IEEE *"A Low-Power Ternary CAM With Positive-Feedback Match-Line Sense Amplifiers"*, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS, VOL. 56, NO. 3, MARCH 2009.

[6] Eun Chu Oh and Paul D. Franzon *"TCAM Core Design in 3D IC for Low Matchline Capacitance and Low Power"* ECE Dept., North Carolina State University, 2410 Campus Shore Drive, Raleigh, NC, USA 27606

[7] Kaustav Banerjee *"Lecture 17 Semiconductor Memory Design-I"* Electrical and Computer Engineering E-mail: kaustav@ece.ucsb.edu

[8] Palanichamy Manikandan Bjørn B. Larsen Einar J. Aas *"Design of Novel CAM Core Cell Structures for an Efficient Implementation of Low Power BCAM System"* Norwegian University of Science and Technology Trondheim, Norway.

[9] Chao-Ching Wang, Jinn-Shyan Wang, Member, IEEE, and Chingwei Yeh *"High-Speed and Low-Power Design Techniques for TCAM Macros"* IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 43, NO. 2, FEBRUARY 2008.

[10] Midas Peng and Sherri Azgomi *"Content Addressable memory (CAM) and its network applications,Altera International Ltd"*.